Can Supervised Machine Learning be used to predict the sales of different vehicle types?

Mark Anthony Micallef, BAN Group
Institute of Information & Communication Technology
Malta College or Arts, Science & Technology
Corradino Hill
Paola PLA 9032

mark.micallef.b42575@mcast.edu.mt

***Abstract*—In this day and age, time consciousness has made the motor vehicle a necessity in everyday life; using these motor vehicles for transportation is convenient and saves a significant amount of energy and time. Motor vehicles have evolved into an essential commodity to improve the quality of living standards. The aim behind this study was to determine if supervised machine learning could be used to make predictions on vehicle sales. The dataset utilized in this study was created by combining news releases from the National Statistics Office (NSO) with the information provided by Transport Malta (TM). This dataset contains numerical values of the stock of licensed motor vehicles by vehicle group and period.**

Simple Linear and Polynomial Regression were chosen as the basis for these predictions and were then compared using all the metrics generated (r2 score, mean square error (MSE), and root mean square error (RMSE)) to get a better understanding of which approach was the best performing for predicting individual vehicle groups. The model with the best overall performance was Simple Linear Regression Model Three. With a 60:40 train/test split, this model had the highest average r2 score and MSE of all the models. The best test case of the Simple Linear Regression was the third iteration of the third model, the average r2 obtained was 0.858551009 (85.8551009%). Even though the highest r2 score was obtained from the third iteration of the third model, the MSE and RMSE highest average were obtained from the second iteration of the second model with 359811.825 and 350.655703, respectively. Another table was created with the highest metrics obtained for each target variable showing the best train/test split and state for the different target variables.

***Index Terms*—Machine Learning, Regression, R2 score, RMSE, MSE**

I. Introduction

The demand for motor vehicles has increased over the years, the main reason that consumers buy motor vehicles is to be able to travel from one place to another with convenience of taking their own vehicle to the destination at any given time. Given that, there is little to no information about how many vehicles are sold from each category of motor vehicles in Malta. With the increase in demand of motor vehicles, this information would be beneficial for car dealers and vehicle importers to improve their conversion rate (sales). The results obtained are generalized over the whole sales in Malta, to be useful to the car dealers and vehicle importers their business data needs to be used to make the predictions on the average number of vehicles they would be selling in the provided time period.

This study stands to fill a literature gap mentioned on the number of motor vehicles under each category that are sold in Malta, having this information would reduce a large number of unnecessary imports which is directly connected to the pollution generated. By reducing the number of imports, the container ships carrying the vehicles would be lighter and the number of trips taken are reduced. This served as a catalyst to pursue this study/research with the aim of coming up with a solution that may outperform or improve present options. The automotive industry is found to be one of the largest industries and will keep expanding over time with new technologies coming out each year, generally the number of sales will increase or decrease over time. In some circumstances, the number of sales for certain categories will spike (either increasing or decreasing) which can be caused by a number of factors including news, laws regulations as well as benefits on certain types of vehicles. In the dataset used, there are some signs of spikes in the data, the reason for these fluctuations in the dataset were not observed in further detail.

The purpose of this study is to evaluate the use of supervised machine learning in predicting the sales of different types of motor vehicles, with the aim of observing how individual target variables perform when predictions are made using historical data. The dependent variables of this study are the different metrics generated included in the r2 score, means square error (MSE) and root mean square error (RMSE). The target variables were utilized as the independent variable. Since the data is separated into distinct time periods, this independent variable in the scatter plot displays the direction of each target variable, which was useful when generating the scatter plots. Given the type of experiment, a control experiment has been performed, with the dataset as the control variable since it has been kept constant throughout the whole experiment. It is hypothesized that it will be possible to document and predict the number of vehicles sold for the different vehicle types. The number of vehicles sold will be predicted using secondary research of motor vehicles and machine learning techniques. To address the hypothesis, three research questions were formulated:

1) What supervised machine learning algorithms can be used to predict the sales of different types of motor vehicles?
2) What characteristics of secondary data are required/searched for in order to achieve reliable results on the number of automobiles sold using the suggested algorithm?
3) What is the proposed experiment's level of accuracy, and would this research be beneficial to any stakeholders?

II. Literature Review

Machine learning aims to answer the question of how to create machines that learn on their own. It is at the interface of computer science and statistics, as well as at the heart of artificial intelligence and data science. This is one of today's fastest expanding technological topics [1]. The development of new learning algorithms and theory, as well as the continual growth in the availability of online data and low-cost processing, have pushed recent advances in machine learning. Data-intensive machine-learning approaches are being used across science, technology, and business, resulting in more evidence-based decision making in a variety of fields, including health care, manufacturing, education, and marketing [2].

A system's capacity to learn is essential in a changing environment. The virtual world has produced a vast amount of data which is accelerating the adoption of machine learning (ML) solutions & practices. Machine learning can be divided into four groups, these include supervised, semi-supervised, unsupervised and reinforcement learning, to support this assertion [3].

Supervised machine learning is the development of algorithms capable of establishing broad patterns and hypotheses by using externally provided cases to predict the outcome of future occurrences. The aim of supervised machine learning classification algorithms is to categorize data based on past knowledge [4]. The task of knowing a function that translates an input to an output based on example input-output pairs is known as supervised learning. It infers a function from a set of training examples and tagged training data. Algorithms that require external aid are known as supervised machine learning algorithms. The training and testing datasets are separated from the input dataset. There is an output variable in the train dataset that has to be predicted or categorised. For prediction or classification, all algorithms learn patterns from the training dataset and apply them to the test dataset [5].

Unsupervised machine learning approaches make it easier to analyze raw information, allowing for the generation of analytic insights from unlabeled data. In order to use any type of machine learning a number of tasks need to be completed. The selection of samples, or objects to cluster, the selection of features to utilize in the clustering, the selection of a similarity measure for sample comparison, and the selection of an algorithm to apply are all necessary for executing unsupervised machine learning [6].

1) Regression

A. Simple Linear Regression

Linear regression is a mathematical approach that uses a linear or straight-line function to try to characterize the connection between two or more variables. The approach may also be used to make inferences about a wider population or data collection, or to make predictions about future data, based on a study of the present data or sample. A variant of linear regression with a single outcome or dependent variable and a single predictor or independent variable is known as simple linear regression [7]. A simple linear regression model builds on the concept of correlation by formalizing a statistical relationship between the two variables in which Y is proportional to X. The covariate, predictor, explanatory, or independent variable is referred to as X. Y, on the other hand, is known as the result, or the expected, response, or dependent variable. A correlation, on the other hand, does not distinguish between which variables are explanatory and which are result [8].

B. Multiple Linear Regression

Multiple linear regression is a variant of simple linear regression in which more than one predictor variable is included. If the investigator thinks that the outcome of interest is linked to or dependent on more than one predictor variable, straightforward linear regression may not be the best technique. A multiple regression model can be used to account for numerous predictor variables at the same time [9] .

C. Polynomial Regression

Polynomial regression analysis is commonly used in curve fitting. In most circumstances, the degree of the polynomial that best describes the data is unknown to the investigator. As a result, in addition to estimating the unknown parameters of the polynomial, the investigator must also determine the degree of the polynomial. It's critical to carefully pick such that the model is smart enough to correctly describe the data while yet being simple enough to use in reality. As the degree of the polynomial increases, it better matches the data. On the other hand, the resulting polynomial isn't especially useful for prediction or interpolation/extrapolation [10].

2) K-means

K-means is a common data mining clustering technique that is commonly used for grouping massive collections of data. It was one of the most basic non-supervised learning methods that was used to tackle the well-known cluster problem. This approach iteratively classifies the provided date objects into k distinct clusters before settling on a local minimum. As a result, the clusters created are compact and autonomous. The algorithm is

divided into two parts. The first step chooses k centers at random, with the value k predetermined. The next step is to transport each data object to the closest data center [11].
Clustering is focused on identifying K components in the data set that may be utilized to generate an initial representation of clusters in this approach. The cluster seeds are made up of these K components. The data set's remaining items are then allocated to one of these clusters. Even though the procedure appears to be simple, it may be difficult to precisely identify the initial K components, or cluster seeds [12].

3) Agglomerative Hierarchical Clustering

Clustering is the process of grouping a collection of things into clusters. Hierarchical clustering is a type of cluster analysis used in data mining that aims to create a hierarchy of clusters. There are two sorts of strategies for hierarchical clustering: Agglomerative: This is a "bottom-up" strategy in which each observation is placed in its own cluster, with pairs of clusters merging as one progresses up the hierarchy. Divisive: This is a "top down" strategy in which all observations begin in one cluster and are divided iteratively as one progresses down the hierarchy. [13]. Cluster analysis is an iterative process of knowledge discovery or participatory multi-objective optimization, not an automated operation. Pre-processing and parameter settings will need to be tweaked until the outcome has the appropriate qualities [13].

Following the research conducted by [14] on Benchmarking of Regression Algorithms and Time Series Analysis techniques for Sales forecasting, the study's goal was to forecast the sale amount of the fashion industry will make. The researchers used a number of regression methods in machine learning and time series analysis approaches to anticipate the volume of sales based on many factors in this study. Walmart sales data was utilized in the Microsoft Azure Machine Learning Studio platform.

The dataset was derived from the Kaggle website for the Walmart Recruiting Store Sales Competition. Features.csv, store.csv, and train.csv are three separate files. Each file contains unique qualities that might be useful to the suggested research. The characteristics. Temperature, gasoline price, and unemployment are all included in the features.csv file. The store.csv file contains information on the store, such as the store id, store type, and store size. Finally, the train.csv file had some historical data as well as actual sales figures. For the sake of this experiment, these three csv files were concatenated into one document.

Because of the time series analytic methodology, the first studies were limited to one Walmart store and one department. Several assessment metrics, such as Mean Absolute Error, Root Mean Squared Error and coefficient of determination (R2 score), were used to assess test findings.

TABLE I
DEPARTMENT RESULTS

Method	RMSE	MAE	R2 score
Bayesian Linear Regression	2469.54	4361.40	0.96
Linear Regression	2480.12	4365.04	0.96
Neural Network Regression	14951.09	22499.33	0.00
Boosted Decision Tree Regression	1669.10	3696.59	0.97

III. RESEARCH METHODOLOGY

The aim of this research as specified also in the title was the forecasting of sales of vehicles with the use of supervised machine learning algorithms. Supervised machine learning was chosen because of its ability to utilize the continuous target variables for forecasting. It makes use of historical data so it can produce accurate results and analysis. R and Python were mainly considered due to their capabilities for the research proposed. Due to a better understanding of the language, the availability of Machine learning (ML) libraries and due to its popularity amongst data scientists in this research area. Python was the chosen language for this experiment, with the use of Microsoft Visual Studio code using the Jupyter notebook framework.

A variety of libraries were used for this research. The Pandas library was mainly used to load, manipulate, and clean the data, matplotlib was used for visualization of the data including histograms, boxplots and other forms of visualization, NumPy was used to make predictions and sklearn was used to import the algorithm libraries as well as for the model selection (train/test split). An environment was established using anaconda prompt to make use of these libraries. Anaconda was installed on the workstation and launched as administrator; visual studio code needed to be closed to create the environment. For the new environment to be created, a series of commands were executed, and this environment was then used throughout the experiment. The commands were as follows:

1) conda create –name environment python 3.8
2) conda activate environment
3) conda install Jupyter
4) ipython kernel install –name "environment-venv" –user
5) conda install -c conda-forge geopandas
6) conda install -c conda-forge matplotlib
7) conda install -c conda-forge pulp
8) conda install -c conda-forge scikit-learn
9) conda install -c conda-forge numpy
10) conda clean -all

The dataset chosen for the research proposed was that of the National Statistics Office (NSO), the premise that the data is posted on the official statistics office website in Malta, with information provided by the primary transportation sector governmental agency with environmental protection responsibilities, was one of the reasons why this dataset was chosen. To download the news releases, the researcher needed

to visit the national statistics office (NSO) Malta website and search for “Motor Vehicles” with the use of the search feature. The article news releases have the title “Motor Vehicles”, they include quarterly information about the motor vehicles.

The individual news releases contained five unique tables; the tables were spread out into three different sheets. Each table contained numerical data on the stock of motor vehicles under different classes (“new”, “used”, “type of fuel used”, etc.). For the scope of this experiment, “Table 1. Stock of licensed motor vehicles by vehicle group and period” was used to predict the sales of motor vehicles.

During the data cleaning process, which was largely focused on accommodating the algorithms that were being tested. The dataset required some minor adjustments. Null or string values are not supported for the simple linear regression model, The data downloaded from the National Statistics Office (NSO) news releases had a string value indicating that the specific category of motor vehicle had no sales in that quarter. For the purpose of this experiment, the string value was replaced with a zero. The news releases issued by the National Statistics Office utilizing information acquired by Transport Malta (TM) have undergone several alterations throughout the years. These developments were brought about by changes in the transportation industry.

These alterations were documented in the news release’s methodological notes page, and they had to be taken into consideration while constructing the dataset. This meant that depending on the information provided in the notes, the columns in the dataset were either combined or modified. An example of this phenomenon was evident in the final quarter of 2013 which mentioned that “Due to the small number of powered bicycles, such data is included with motorcycles”. These changes were acknowledged in the methodological comments and needed to be corrected in the past data. For this experiment, the day and month taken corresponded to the final day of the quarter using the format (dd/mm/yyyy). The year is taken from the “Year” column in the dataset and placed at the end of the column (ex. for the first quarter of 2014, the date format would be 31/03/2014). After the news releases from the years 2010 to 2020 were combined into a single document and was cleaned, the document was saved as a .csv file with the name “data” and imported into the visual studio code using the pandas library.

After the data was cleaned and prepared for the experiment, the implementation stage was initialised. In this stage, a number of steps were followed to reach the goal of answering the hypothesis and research questions stated at the start of the research. These steps include loading the data, visualisation, creating the model and making a number of iterations. The results obtained are then documented and displayed to be later individually discussed and compared in the results and findings section of the study.

After manually evaluating and modifying the dataset to meet the demands of the research the data was loaded into a data frame using Python’s pandas package library.

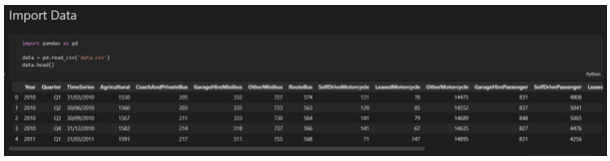

Fig. 1. Using the pandas library to import the .csv dataset

The data frame was then displayed in a table with a variety of properties from which data exploration can begin, this includes interpreting metrics like mean, standard deviation, count, and many others. This was done to better visualize the dataset and it might also disclose additional information that was missed while manually exploring the data.

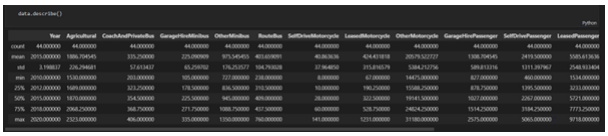

Fig. 2. Using the .describe() to explore the data and interpret the metrics

The data was then visualized using the matplotlib library after it had been explored. Using the scatter plot, these graphs were plotted to get a visual idea of how the data was distributed. This was helpful when determining which algorithm was the best fit for each individual target. This still needed to be validated to confirm that the assumptions made were correct, which could be done using the algorithm’s metrics. Both the ”Agricultural” and the ”RouteBus” target variables provide an example of this. Since the data in the Agricultural scatter plot follows a straight line, a linear regression would be more appropriate than a polynomial regression or a clustering technique.

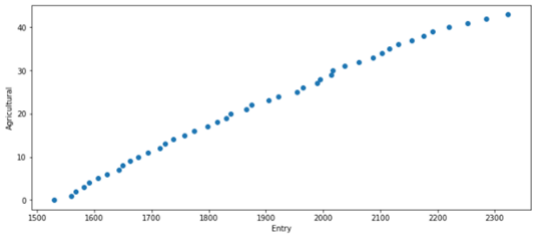

Fig. 3. Agricultural target variable scatter plot

The scatter plot of the ”RouteBus” target variable did not follow a straight line; rather, the graph indicated that the data was grouped into a small number of clusters, The number of cluster groups identified from the graphs were 3/4 groups. This target variable was evaluated using hierarchical clustering based on this information.

For each target variable, boxplots were also plotted. The boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their

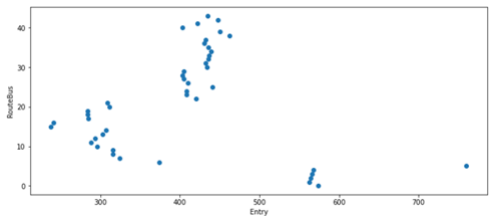

Fig. 4. RouteBus target variable scatter plot

quartiles. In addition, the boxplot identifies outliers in the targets being tested.

After the visualisation stage of the research, the model creation stage was prepared. For each algorithm, a number of steps were taken. Since both simple linear regression and polynomial regression are regression methods, the steps for creating the model for the experiment are similar to each other. Agglomerative Hierarchical clustering incorporates some of the phases seen in the development of the regression model, as well as elements that were unique to agglomerative clustering.

Starting with the Regression analysis, the indexing of the target variables, stored in Variable X in both simple and polynomial regression had to be reshaped from a 1-dimensional array to a 2-dimensional array in order to be used in the regression model.

```
# Reshape X since it is a 1D array
X = X.reshape(-1,1)
```

Fig. 5. Reshaping Variable X from a 1-dimensional array to a 2-dimensional array

Data splitting is then required, which is the act of dividing data into two independent sets: training and testing. To ensure that the model is not evaluated using the same data as it was trained with. The training set, also known as in-sample data or training data is a collection of data that was used to analyse and build the model's rule. The test set is a set of data that wasn't utilized to train the model or influence the validation set's input characteristics. The train/test split was imported from the sklearn library. Three models were built for this study, each with a distinct train/test split (70:30, 60:40, and 80:20, respectively). Each model went through three iterations, each with a distinct seed to ensure full randomness.

```
from sklearn.model_selection import train_test_split

state = 0
# in this case we will be using 70% of the data to train the model,
# and 30% of it to test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=state)
state += 1

print ('Training set {:} rows'.format(X_train.shape[0]))
print ('Test set {:} rows'.format(X_test.shape[0]))
```

Fig. 6. Train/test split imported using the sklearn.model_selection

The linear regression model was also imported from the same sklearn library as the train/test split. The modthe results.el was fitted using the variables generated in the train/test split.

```
from sklearn.linear_model import LinearRegression
model = LinearRegression().fit(X_train, y_train)
print(model)

LinearRegression()
```

Fig. 7. The linear regression model imported and fitted using the X_train and y_train variables

After the environment has been set and the model was created, the predictions were generated using the Numpy library which was initially imported into the Anaconda environment. The results generated were stores in a variable under the name "predictions". This variable was set to be later used in the experiment.

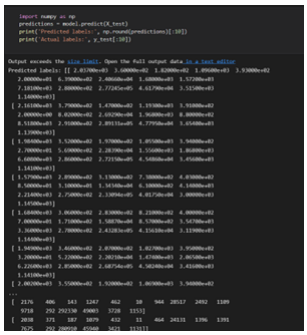

Fig. 8. Import of the numpy library and the predictions that were generated

```
#Agricultural

ax[i].scatter(X_test, y_test[:, i], color="blue")
ax[i].scatter(X_test, predictions[:, i], color="red")
ax[i].set_xlabel('Actual Agricultural')
ax[i].set_ylabel('Predicted Agricultural')
ax[i].set_title("Agricultural Predictions")
ax[i].grid()

polynomial = np.polyfit(X_train[:,0], y_train[:,i], 1)
modelpoly = np.poly1d(polynomial)
rng = list(range(2,45,1))
ax[i].plot(rng, modelpoly(rng), color="green")

r2 = r2_score(y_test[:, i], predictions[:, i])
print ('Agricultural R2: ', r2)

mse = mean_squared_error(y_test[:, i], predictions[:,i])
print ('MSE: ', mse)
rmse = np.sqrt(mse)
print ('RMSE: ', rmse)
print ("")

i += 1
```

Fig. 9. Agricultural data was plotted (simple linear regression), and metrics were generated. Model 1 – 70:30 train/test split

It was clear that neither simple linear regression nor polynomial regression were the best options in some circumstances; this was first apparent during the data visualisation portion of the study and was subsequently verified when the metrics for simple and polynomial regression were provided. The same number of models, iterations, and seeds were utilized for agglomerative clustering. The method might have been created using a loop to speed up the research; however, because the number of clusters would vary based on the target variable, this was not done. Since the output is a percentage per cluster

number, the accuracy was created and then eliminated because it would always result in 100 percent accuracy (if 3 clusters are set, it would be split into 25 percent , 50 percent and 25 percent which add up to 100 percent for example). Argmax was also applied; it took the cluster with the largest percentage of clusters (cluster 2 of 50%, for example) and outputted that percentage. As a result, to choose the optimal output, a visual comparison is employed.

Fig. 10. Agglomerative Clustering for the 4 target variables with low r2 score for Model 2 Iteration 2

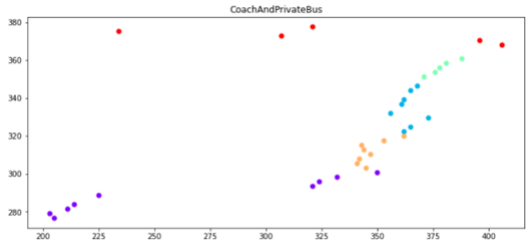

Fig. 11. Agglomerative Clustering Output for Model 2 Iteration 2 for the "CoachAndPrivateBus" target variable.

A total of three model and three iterations with various random states were established during the experiment. Each model was iterated three times in order to give enough test cases to compare and contrast the findings. As a result, each repetition will be treated as a test case. The metrics of each target variable (R2 score, Root Mean Square Error (RMSE), and Mean Square Error (MSE)) will be included in each iteration's output, as well as a scatter plot for each target variable with the projected value, actual value, and the line of best fit to evaluate the algorithm's accuracy.

	Target Variables – Motor Vehicles	R2 score	MSE	RMSE
Model: 3 Iteration: 2 Random State: 444	Agricultural	0.9942996025536305	277.76847413975594	16.666387555188916
	CoachAndPrivateBus	0.6768621302813463	2342.8513225756396	48.403009437179
	GarageHireMinibus	0.9895745383238198	43.06774307105463	6.562601852242343
	Other Minibus	0.9659893349252909	932.7623292285225	30.541157954938814
	RouteBus	0.6803122470930842	18100.23074774928	134.53709803526044
	SelfDriveMotorcycle	0.8013505497962173	307.9783823395478	17.549312873715 25
	LeasedMotorcycle	0.813232146745297	14496.66079338987	120.40207968880716
	OtherMotorcycle	0.9345365455365706	1694532.4511827645	1301.7420832034143
Train/Test Split: 60:40 Training Set: 26 Test Set: 18	GarageHirePassenger	0.7104212593050687	75162.53687179544	274.1578685206672
	SelfDrivePassenger	0.9384552233412639	112004.50283290532	334.67073793940415
	LeasedPassenger	0.9599444083778301	274482.47359367035	523.9107496450806
	TaxiPassenger	0.678510413994718	271.8737382187716	16.488594185641528
	OtherPassenger	0.9913256198896143	4114688.8293141415	2028.4695781091077
	GoodsCommercial	0.8630860188191389	1107814.9580225113	1052.5278894274068
	SpecialCommercial	0.8653986440983472	8641.812930135235	92.96135180888471
	RoadTractor	0.679061544464304 7	1625.2541528540107	40.31444099642225

Fig. 12. Metrics generated for Model three Iteration two

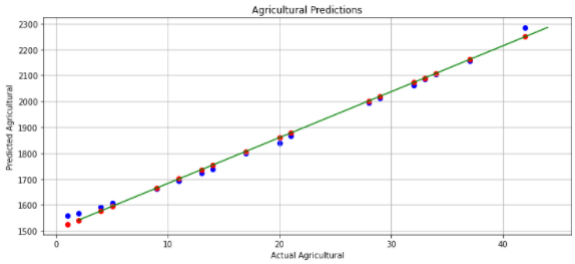

Fig. 13. Scatter Plot of the "Agricultural" target variable for the second iteration of the third model

IV. Findings & Discussion of Results

The findings of the approaches outlined in the previous chapter are presented in this chapter. A full evaluation of the models that were examined on the motor vehicles dataset presented in the methodology section. The metrics generated were compared to determine the best trustworthy model for each target variable to be used in the proposed sales forecasting system.

Scatter plots were created for the visualisation as mentioned in the methodology to have a better understanding of the results gathered. For each of the target variables, a scatter plot was created with the predicted values, actual values, and the line of best fit. The predicted value, which was calculated using the X test data and the predictions, illustrates where the real value should be. The real value that corresponded to actual data was created by repeating the X test and the y test. The line of best fit, also known as a trendline, was an estimate as to where a linear equation may appear in a collection of data presented on a scatter plot; these points corresponded to the projected value's points.

To help understand the visualisation result, metrics were generated for each target variable. These metrics include the r2 score, mean squared error (MSE), and root mean squared error (RMSE). The r2 score or r-squared value discusses that, a linear model explains the percentage response variation in the target variable. This indicates that a high r2 value suggests that the model is better fitted to the data, resulting in more accurate results. The mean squared error (MSE) measures the distance between a regression line and a set of points. It achieves it by squaring the distances (errors) between the points and the regression line. Squaring is required to eliminate any negative signals. It also gives larger discrepancies more weight. Since the average of a set of errors can be established, it is referred to as the mean squared error; the smaller the mean squared error, the more accurate the forecast. The root mean squared error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. This indicates how densely the data is concentrated around the line of best fit. RMSE is widely used to check experimental results. This information was relevant when using either simple linear regression or polynomial regression since they are both types of regression.

Clusters aims to uncover structure in data by grouping data that has similar qualities together. Hierarchical clustering is a type of clustering technique that generates 1-to-n groups, with n representing the number of observations in the data collection. The clusters grow increasingly similar as you progress down the hierarchy from one cluster (which includes all of the data) to n clusters (each observation being its own cluster). Agglomerative hierarchical clustering was employed in this study. This was appropriate since the dataset utilized had a modest number of rows, it was ideal since the number of clusters needed to be identified for each target variable in each iteration and model being tested with this clustering method.

In the methodology section, the test case result of second iteration of the third model was displayed along with the different metrics acquired. The metrics obtained for this iteration was addressed in further depth in this section, with the goal of determining the best method. The r2 value, which ranges from 0 to 1, was a decisive factor in simple linear regression and polynomial regression. This value indicated how well the model depicts the target variable. The mean square error (MSE) and the root mean square error (RMSE) were two other important factors to consider. The scatter plot graph was used to evaluate the agglomerative clustering outputs generated. The scatter plots created with the number of estimators for each iteration are reviewed visually against the target variables being tested to find the optimal iteration for the target variables being tested.

Model Three - Test Case Two - 60:40 train/test split: Random State 444

The metrics acquired during the second iteration of this model were better than those obtained during the first iteration. Given that, some of the metrics acquired in certain target variables tested were inferior than those produced in the model's third iteration. The best generated metrics were those of the "Agricultural" target variable, the data points chosen were the result of an r2 score of 0.9942996025536305 (99.42996025536305%) which meant that this iteration explains 99.42996025536305% variation in the target variable.

Some of the target variables didn't achieve satisfactory results with the use of simple linear regression. These target variables were then tested using Simple linear regression. After testing both algorithms, the results obtained from the polynomial regression for the target variables in question was better than those obtained when using simple linear regression. The improvements were recorded showing the results of using one algorithm over the other.

	Target Variables – Motor Vehicles	Using Polynomial Regression	Using Simple Linear Regression
Model: 3 Iteration: 2 Random: 444 Split: 60:40 Training: 26 Test: 18	CoachAndPrivateBus	0.6768621302813463	0.4189072156905994
	RouteBus	0.6803122470930842	0.17132069743679823
	TaxiPassenger	0.678510413994718	0.14091547180617525
	RoadTractor	0.6790615444643047	0.0007185626171586357

Fig. 14. Table of the r2 scores obtained from both simple linear and polynomial regression for the target variables mentioned

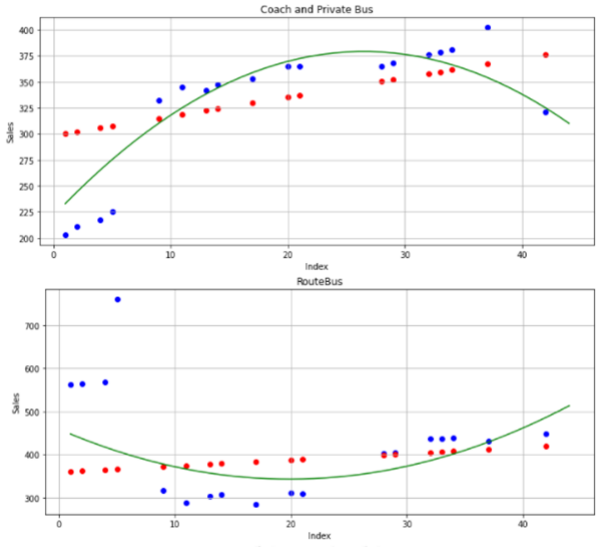

Fig. 15. Polynomial Regression plotted for the CoachAndPrivateBus and RouteBus target variables

Agglomerative Hierarchical Clustering was then tested to determine if the results would be superior than those obtained by Simple and Polynomial Regression. The actual train data was shown on the y-axis while the forecasts were plotted on the x-axis. This meant that by comparing the actual data value

with the projected value, the data could be noticed. For high accuracy, the projected value must match the observed data exactly. To evaluate each cluster's correctness, this had to be shown for each cluster.

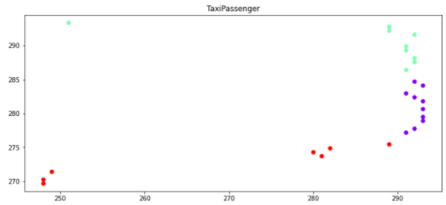

Fig. 16. TaxiPassenger Agglomerative Clustering output

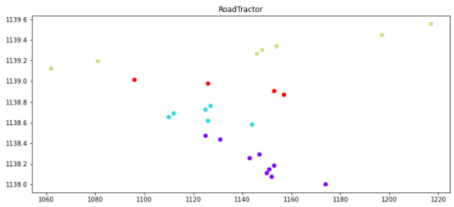

Fig. 160 – RoadTractor Agglomerative Clustering – Model 3 – Iteration 2

Fig. 17. RoadTractor Agglomerative Clustering output

The research conducted by [14] on Benchmarking of Regression Algorithms and Time Series Analysis techniques for Sales forecasting was found and mentioned in chapter 2. The researcher made use of a number of algorithms to forecast the sales amount of the fashion industry. This industry was difficult to predict since a number of factors affect the sales volume, this concept was similar to motor vehicles. Through unpredictable factors, the sales of motor vehicles can increase or decrease. In the first investigation, the dataset was cleaned and reduced to one department within a Walmart store, and RMSE and MSE were calculated for each of the methods. After that, the R2 score was calculated on the entire dataset and across all departments.

The metrics generated by the linear regression algorithm for the research done by [14] was that of a 96% R2 score, an RMSE score of 2480.12 and an MAE score of 4365.04. This exceeds the results obtained from this research for the R2 score since a high number of target variables are being tested and the average is taken from all the target variables. The highest R2 score of the whole iteration is 0.858551009 (85.8551009%), the lowest MSE score was 359,811.825 and the lowest RMSE was that of 350.655703. The RMSE score of this research were found lower than that of the other researcher even though a lower R2 score is obtained.

Simple Linear Regression was clearly the highest performing method among all the simple linear regression, polynomial regression, and agglomerative clustering models created and trained, based on the findings acquired throughout these trials. Although this method did not represent all of the target variables, the metrics obtained from the test cases with the use of this algorithm showed substantial results.

From all the models generated, the best average r2 score of 0.858551009 (85.86% taken to 2 decimal places) was generated by the third iteration of the third model. This meant that this model explains the percentage response variation in the target variables. Even though the highest r2 score was generated from the third iteration of the third model, The second iteration of the second model produced the best average MSE and RMSE result (lowest value). The best average output of the mean square error (MSE) and root mean squared error (RMSE) obtained from the experiment as shown in table 34 were 359,811.825 and 350.655703 respectively. The MSE score represents the average distance between the line of best fit and a set of points.

The individual outputs were displayed in the methodology section showing the MSE of the target variables being used for this experiment. These outputs were further discussed in regression analysis, on how some of these outputs were obtained and visualised through scatterplots. The RMSE is a measurement of how far the regression line data points are separated by the standard deviation of the errors/residuals. This reflected how concentrated the data was around the line of best fit.

The performance of the other two algorithms was mainly observed in the target variables which didn't achieve satisfactory results with the use of simple linear regression. There were mainly four target variables which both polynomial regression and agglomerative clustering were used to attempt to get a better representation. Some improvements were discovered when compared to simple linear regression, this was not evident in all the models/iterations due to the different data selected during the train/test split. The metrics used for the polynomial regression were the same as those of the simple linear regression, this made it easier to determine if the results improved or deteriorated with the change of algorithms used. As for the agglomerative regression, it was slightly more difficult to determine the model accuracy mainly due to the fact that visual comparison was used to compare the different models to the algorithm's metrics.

V. Conclusion

In this study, several target variables were studied using various techniques. The goal was to determine which algorithm would produce the best metrics for each of the target variables studied. These variables had to be examined and assessed separately since they don't follow the same set of patterns, which meant that just because an algorithm achieved satisfactory metrics for one target variable it didn't guarantee the same results can be reached for the others.

The scatter plot graphs created during the prototype's visualisation stage were primarily used to identify which algorithm would best match the data plotted, and this selection was later put to the test to validate if the algorithm chosen was the best choice. When the results obtained were interpreted in the results and findings, the r2 scores obtained from both the simple linear and polynomial regression were documented in

a table. These tables were used to justify why the change in the algorithm was beneficial for the results obtained.

TABLE II
AVERAGE SCORES OF EACH MODEL

Model	Train/Test	R2 score	MSE	RMSE
1	70:30	0.840728809	499,863.6434	394.1954477
2	80:20	0.838392409	463,658.8383	371.3991513
3	60:40	0.851500477	461,464.628	375.8254266

Taking the results from the table shown above, the best model with the given dataset was the third model meaning the 60:40 train/test split. This train/test split achieved the highest average r2 score and best average mean squared error (MSE) while the second model (80:20 train/test split) achieved the best average root mean square error (RMSE).

In this study, some limitations were discovered, mainly in the research stage of the experiment, the first limitation was the the amount of research articles identified that utilize simple linear regression for predictions on motor vehicles was the first restriction discovered in this study; the majority of research papers found employ multiple linear regression to produce these predictions. The second limitation was on the news releases found on the National Statistics Office (NSO) Malta website, the data provided was divided into quarters, this had an effect on determining what was the cause of any spike and/or any abnormalities in the data. Given that this time period was provided, a number of factors could have affected a single numerical value such as a change in the law, benefits on a form of transportation or the change in registration.

The experiment only achieved a TRL level of three, indicating that the study remained in the research phase, with basic principles observed, a technological idea established, and an experiment proof of concept confirmed to make analytical predictions. The improvement would be to take the experiment to higher levels, this means taking the prototype and testing its range in a real-world environment. This would provide evidence of the experiment performance in an outside environment.

In this research study, it was hypothesized that it would be possible to document and predict the number of vehicles sold for the different vehicle types. The number of vehicles sold would be predicted using secondary research of motor vehicles and machine learning techniques.

1) What supervised machine learning algorithms can be used to predict the sales of different types of motor vehicles?

A. In this study, three different machine learning algorithms were used to establish which algorithm generates the best metrics for the dataset used, the simple linear regression which achieved the best metrics from the three algorithms tested, achieving accuracies of over 90% for some of the target variables. Polynomial Regression was used on target variables that achieved low accuracy for table variables when using simple linear.

2) What characteristics of secondary data are required/searched for in order to achieve reliable results on the number of automobiles sold using the suggested algorithm?

B. During the research phase, three datasets were discovered: a Kaggle dataset, an SMMT dataset, and news releases from the National Statistics Office (NSO). For this experiment, the NSO's news releases were used, and a dataset was created to fit the algorithms. The location the data was based on, where the data was released, and who provided the information were the characteristics that led to this choice on the dataset. Accurate findings for a variety of motor vehicles were achieved using secondary data.

3) What is the proposed experiment's level of accuracy, and would this research be beneficial to any stakeholders?

C. The suggested experiment's accuracy may be evaluated in two ways: the first is by looking at the accuracy achieved from the entire iteration or model, and the second is by looking at the results produced for each of the target variables (motor vehicles). If the first strategy is used, the third model (60:40 train/test split) had the best level of accuracy for the model, with an average r2 score of 0.851500477. (85.1500477%). For the outcomes of the second method, a table was made to track where the best metrics for each target variable were reached

APPENDIX

https://nso.gov.mt/en/News_Releases/View_by_Unit/Unit_02/Regional_and_Geospatial_Statistics/Pages/Motor-Vehicles.aspx

REFERENCES

[1] M. Jordan and T. Mitchell, "Machine learning trends, perspectives, and prospects." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26185243/

[2] A. C. Muller and S. Guido, "Introduction to machine learning with python." [Online]. Available: https://www.abebooks.it/9781449369415/Introduction-Machine-Learning-Python-Guide-1449369413/plp

[3] M. G. Pecht and M. Kang, "Prognostics and health management of electronics." [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119515326

[4] l. G. Maglogiannis, "Emerging artificial intelligence applications in computer engineering." [Online]. Available: https://books.google.com.mt/books/about/Emerging_Artificial_Intelligence_Applica.html?id=e52lAQAACAAJ&redir_esc=y

[5] M. Praveena and V. Jaignesh, "A literature review on supervised machine learning algorithms and boosting process. international journal of computer applications." [Online]. Available: https://www.researchgate.net/publication/318486479_A_Literature_Review_on_Supervised_Machine_Learning_Algorithms_and_Boosting_Process

[6] R. Gentleman and J. Carey, "Unupervised machine learning." [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-77240-0_10

[7] K. Marill, "Advanced statistics: Linear regression part ii: Multiple linear regression." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/14709437/

[8] L. Eberly, "Multiple linear regression: In topics in biostatistics." [Online]. Available: https://experts.umn.edu/en/publications/multiple-linear-regression

[9] K. Marill, "Advanced statistics: Linear regression part i: Simple linear regression." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/14709437/

[10] S. Narula, "Orthogonal polynomial regression." [Online]. Available: https://www.jstor.org/stable/1403204

[11] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm." [Online]. Available: https://www.semanticscholar.org/paper/Research-on-k-means-Clustering-Algorithm:-An-Na-Xumin/024abd649e2393e57951e9eaadee8372cc054658

[12] S. Bhatia, "May adaptive k-means clustering." [Online]. Available: https://www.aaai.org/Library/FLAIRS/2004/flairs04-119.php

[13] K. Sasirekha and P. Baby, "Agglomerative hierarchical clustering algorithm." [Online]. Available: https://www.semanticscholar.org/paper/Agglomerative-Hierarchical-Clustering-Algorithm-A-Sasirekha-Baby/bd9fe11a960001cb845ea74a75cf8c10b8c34615

[14] C. Cagatay, E. KAAN, A. Begum, and A. Akhan, "Benchmarking of regression algorithms and time series analysis." [Online]. Available: https://www.researchgate.net/publication/330762801_Benchmarking_of_Regression_Algorithms_and_Time_Series_Analysis_Techniques_for_Sales_Forecasting